Clustering – K-means

Prerequisite

- Vector algebra
- Distance metrics

Objectives

- Understanding unsupervised learning
- What is clustering
- Different types of distance metrics
- K-means clustering
- Selections of optimal k value –(Elbow method and Silhouette)

Unsupervised learning-

Unsupervised learning is the different type of machine learning algorithm specially used when the target variable is absent in dataset or the dataset is not labelled. The primary goal of unsupervised learning is find hidden pattern exist within the data. This can be achieved through grouping the data points into homogenous groups also known as clusters. Unsupervised learning is much harder and complex than supervised learning as the results are hard to define in the absence of target variable or labels. The other issue with unsupervised learning is the definition of objective function. Despite these issues unsupervised (clustering) is used to gain the insight of data before applying a classification model.

Clustering-

Clustering is one of fundamental problem of unsupervised learning algorithm defined as the process of dividing the simple data points into homogeneous groups of similar data points. Grouping of data points is based on the similarities and dissimilarities between them, the points within same group are similar and among different groups they are dissimilar. This could be understood by following figure where similar data points are clustered on the basis of colour.

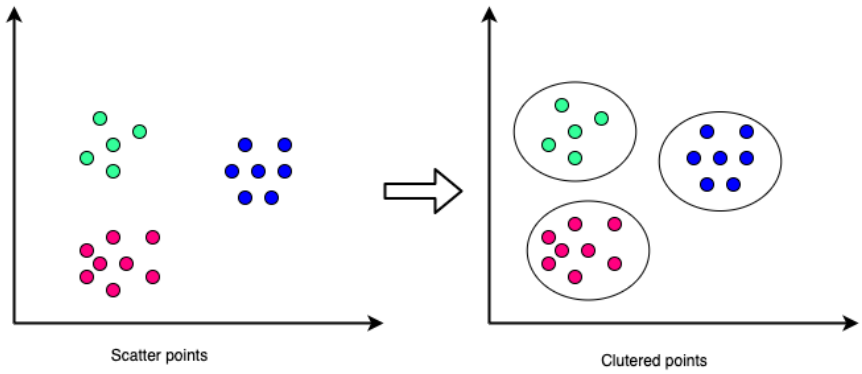


Figure 1 Clustering

Types of clustering-

Primarily clustering is classified as Partitional clustering and Hierarchical clustering. In this note we will describe partitional clustering. Partitional clustering is the type of clustering in which the data points within the dataset are classified into different groups based on their similarities. Ex: K-means Clustering, CLARA etc.

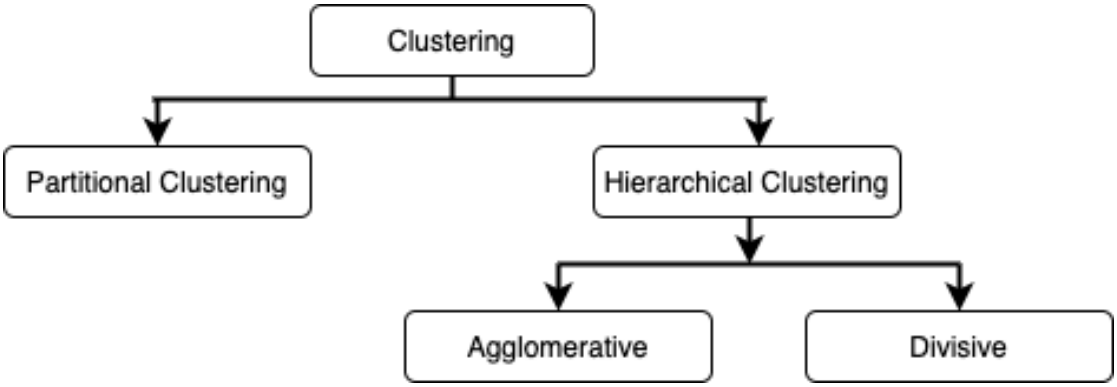


Figure 2 Types of clustering

K-means Clustering:

K-means clustering is a unsupervised learning algorithm whose goal is to find groups or assign the data points to clusters on the basis of their similarity. Which means the points in same cluster are similar to each other and in different clusters are dissimilar with each other. It was developed by researcher named James Macqueen in 1967. Here K means the number of clusters.

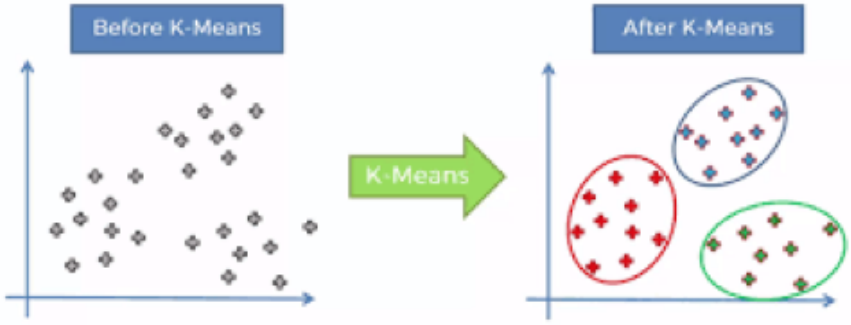


Figure 3 K means Clustering

The above figure is showing the working of K-means clustering. Left part is showing the data points before applying the K-means whereas the right part is depicting the situation after the clustering. (data points are grouped into clusters. K-means clustering is very useful and beneficial when most of the data is in unorganized manner. Before moving forward let define some terminology.

Cluster- It is a collection of the given data points accumulated together because of certain similarities between them.

Centroid- It can be termed as a real or imaginary location which represents the centre of the cluster.

Parameter K- K is a target variable which refers to the number of centroids in the respective or given dataset. It is used to label the new data.

Means- In K-means clustering, **'means'** refer to the averaging of data used to find the centroid in a the cluster.

The working of K-means clustering is depends on the distance metrics, which are used to find the similarity within data points. The popular distance metric are:

Euclidean Distance-

It is the most commonly used distance metrics and defined as the square root of the sum of squared differences between the two points. Let the two points are P(x_1, x_2) and Q (y_1, y_2) the Euclidean distance is given by:

$$PQ_{Euclidean} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

In general

$$PQ_{Euclidean} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan Distance-

It is also known as city block distance or absolute distance. The distance measure is inspired with the structure of Manhattan city where the distance between two points is measured through city road grids. The distance is defined as the sum of absolute differences between two points coordinates.

$$PQ_{Manhattan} = |x_1 - y_1| + |x_2 - y_2|$$

or

$$PQ_{Manhattan} = \sum_{i=1}^{n}|x_i - y_i|$$

Chebyshev Distance-

This distance is also known as Maximum value distance or chessboard distance. The distance is based on absolute magnitude between the coordinates of pair of two points. This distance is equally used with the quantitative and ordinal variable.

$$PQ_{Chebyshev} = Max(|x_1 - y_1|, |x_2 - y_2|)$$

or

$$PQ_{Chebyshev} = Max(|x_i - y_i|)$$

Minkowski Distance-

Minkowski distance is one of the generalized distance measures, which means that by manipulating the formula different distances measures can be obtained. Above stated distance measures are the special case of Minkowski distance.

$$PQ_{Minkowski} = (\sum_{i}^{n}(x_i - y_i)^p)^{\frac{1}{p}}$$

When p =1 Minkowski has become Manhattan distance.

When p =2 Minkowski has become Euclidean distance.

When p = ∞ Minkowski has become Chebyshev distance.

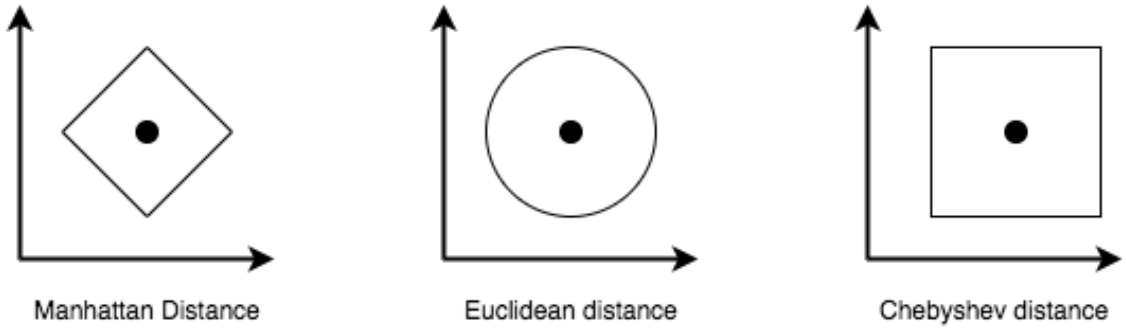


Figure: Distance metrics

Mahalanobis Distance-

This distance measure is used to calculate the distance between two points in multivariate space. The idea is to calculate the distance of a point P from any distribution D in terms of standard deviation and mean of distribution D. The main advantage of mahalanobis distance is that it includes the covariance of distribution to measure the similarity between two points. The distance equation is given by:

$$PQ_{mahalaobis} = \sqrt{(P-Q)^T S^{-1}(P-Q)}$$

Where P and Q are two random vectors of same distribution and S is covariance matrix.

NOTE: Most widely used distance measure is Euclidean distance, But all the distances have their respective purpose and importance. One cannot say or claim that only one particular distance measure is always accurate.

Working of K-means Clustering- The working of k-means clustering can be summarized as:

Step 1- Initialize the K random centroids or k points.(There can be two strategy for it.)

i. Pick random data points and consider those as starting points.

ii. Choose K random values for each particular variable.

Step 2- For each data point calculate the distance of it from randomly chosen K centroid C_i and assign each point to minimum distance cluster.

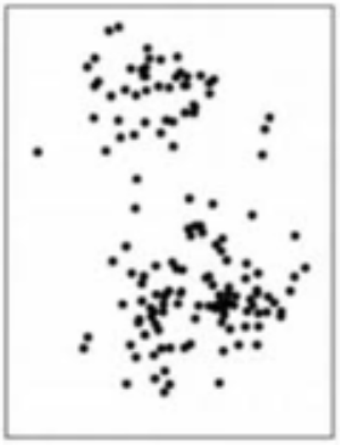
Step 3- Update the centroid by using newly assigned data points to the cluster by calculating the average of data points.

$$V_i = \frac{1}{|C_i|}\sum_{j=1}^{C_i} x_j$$

Where x_j data points within cluster C_i and V_i is vector for centroid C_i.

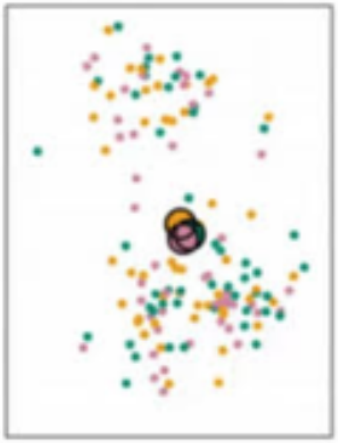
Step 4- Repeat the above process for a given no. of iterations or until the centroid allocation no longer changes.

The algorithm is said to be converged once there are no more changes in the values of centroids. Following figure is showing the above stated process of k-means clustering.

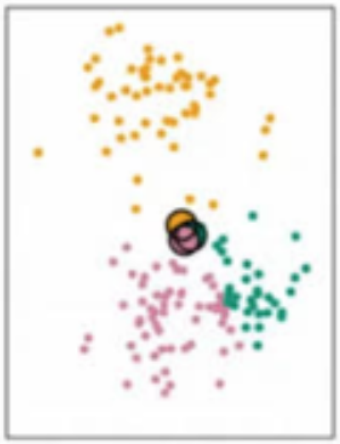


(a) Raw data

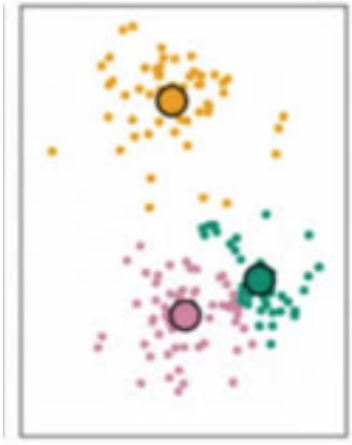
(b) Choosing number of Cluster



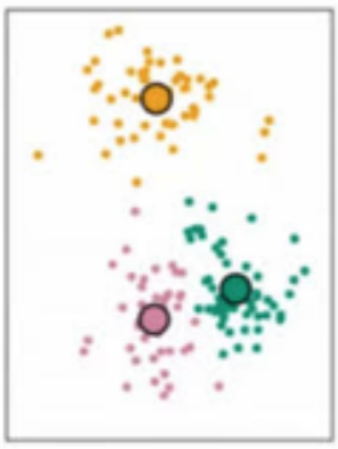
(c) Assigning Random centroid



(d) Calculating similarity



(e) Updating Centroid



(f) Final Result

Objective of clustering-

The objective of clustering is to minimize the distance between data points and its centroid, which can also be expressed as square error term defined as:

$$\min J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^j - C_j\right\|^2$$

Choosing No. of Clusters(value of k)-

Basically, there is no such method for define the exact value of **k**, but there is quick rule of thumb which can be used as heuristic to get the maximum number of clusters as $k = \sqrt{\frac{N}{2}}$.

where N= no. of data points. For example, take N=200 data points, then no. of clusters will be:

$$k = \sqrt{\frac{200}{2}} = \sqrt{100} = 10$$

Hence 10 clusters will be formed for 200 data points.

There are some other techniques from which can be used to find the approximate or optimal value of **k**. These techniques includes:

1. Elbow Method
2. Silhouette method
3. The information theoretic jump, the information criteria.

This note will discuss the Elbow method (which is most popular) and Silhouette method to determine the optimal value of number of clusters.

Elbow method-

It is most popular and well-known method to find the optimal no. of clusters or the value of k in the process of clustering. This method is based of plotting the value of cost function against different values of **k**. As the number of clusters (**k**) increase lesser number of points fall within clusters or around the centroids. Hence the average distortion decreases with the increase of number of clusters. The point where the distortion declines most is said to be the **elbow point** and define the optimal number of clusters for dataset.

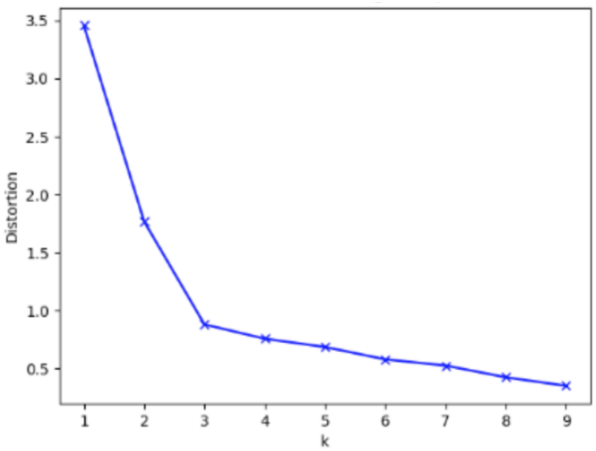


Figure 4 Elbow method

As it is clear from above figure, the distortion declines most at 3. Hence the optimal value of k will be 3 for performing the clustering. In other words the plot looks as an arm with an elbow at k = 3.

Silhouette Method-

Silhouette is a different method to determine optimal number of clusters for given dataset. It defines as a coefficient of measure of how similar an observation to its own cluster compared to that of other clusters. The range of silhouette coefficient varies between -1 to 1. 1 value indicate that an observation is far from its neighbouring cluster and close to its own whereas -1 denotes that an observation is close to neighbouring cluster than its own cluster. The 0 value indicate the presence of observation on boundary of two clusters. Silhouette coefficient is defined as:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & a(i) < b(i) \\ 0 & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & a(i) > b(i) \end{cases}$$

Where $a(i)$ is the distance of observation within its own cluster and $b(i)$ is the distance of observation with its neighbouring cluster.

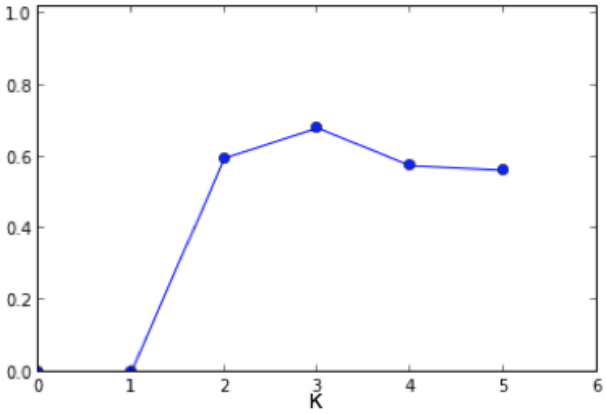


Figure 5 Silhouette coefficient

Advantages of K-means clustering:

- Ease of implementation.
- It works great on large scale data.
- Results guarantees convergence.
- Easily works with new examples.

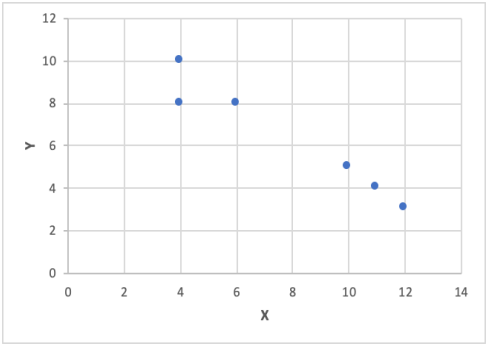
Disadvantages of K-means clustering:

- It is quite difficult to predict number of clusters.
- Initialization of the cluster center is a really crucial part and also somewhat tough.
- K- means clustering only handles numeric data.

Example-

Let we have given with following dataset for clustering.

X	Y
11	4
12	3
10	5
4	10
4	8
6	8



Step 1- Let choose two random centers for clustering as C1(6, 8) and C2(10, 5).

Step 2- For each data point calculate the distance from each cluster centroid C_i and assign each point to minimum distance cluster as:

X	Y	C1(6,8)	C2(10,5)	Label
11	4	6.40	1.41	C2
12	3	7.81	2.83	C2
10	5	5.00	0.00	C2
4	10	2.83	7.81	C1
4	8	2.00	6.71	C1
6	8	0.00	5.00	C1

Step 3- Update the centroid by using newly assigned data points to the cluster by calculating the average of data points.

$$C1 = \left(\frac{4+4+6}{3}, \frac{10+8+8}{3}\right) = (4.66, 8.66)$$

$$C2 = \left(\frac{11+12+10}{3}, \frac{4+3+5}{3}\right) = (11, 4)$$

Step 4- Repeat above steps with new centroids C1(4.66, 8.66) and C2(11, 4).

****